UNIVERSITY of ROCHESTER

air AUDIO INFORMATION RESEARCH

Mitigating Cross-Database Differences for Learning Unified HRTF Representation

***Yutong Wen**, You Zhang, Zhiyao Duan*

University of Rochester, NY, USA

ywen6@u.rochester.edu

IEEE WASPAA 2023

Head-related transfer functions (HRTFs)

HRTFs encode **spectral changes** of sound from source to listener's ears.

HRTFs are fundamental to virtual auditory displays.

It is **hard to measure** personalized HRTFs.

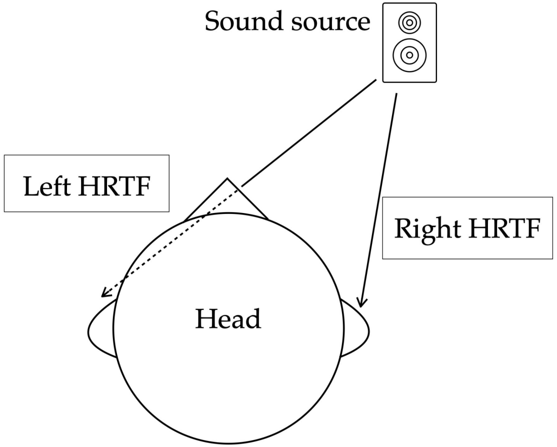
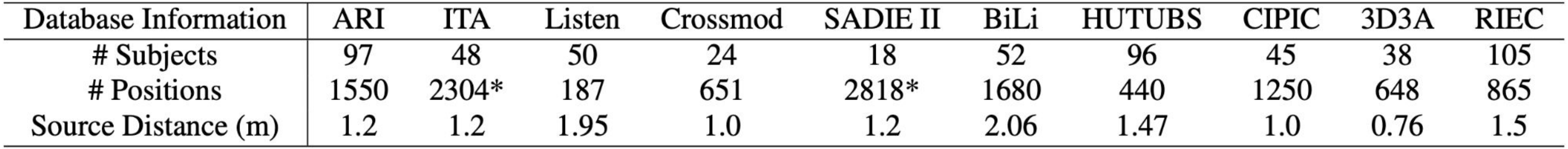


Figure from: https://ieeexplore.ieee.org/document/7099223

Existing databases



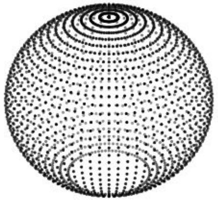
Database Information	ARI	ITA	Listen	Crossmod	SADIE II	BiLi	HUTUBS	CIPIC	3D3A	RIEC
# Subjects	97	48	50	24	18	52	96	45	38	105
# Positions	1550	2304*	187	651	2818*	1680	440	1250	648	865
Source Distance (m)	1.2	1.2	1.95	1.0	1.2	2.06	1.47	1.0	0.76	1.5

Small number of subjects in each database

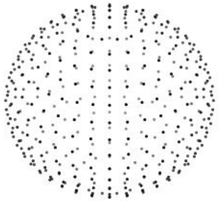
- Hard to model such high-dimensional data

Different spatial sampling schemes

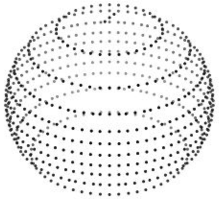
- Hard to do mix-database training



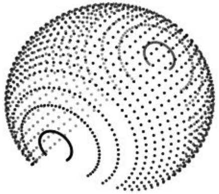
ITA



HUTUBS



Crossmod

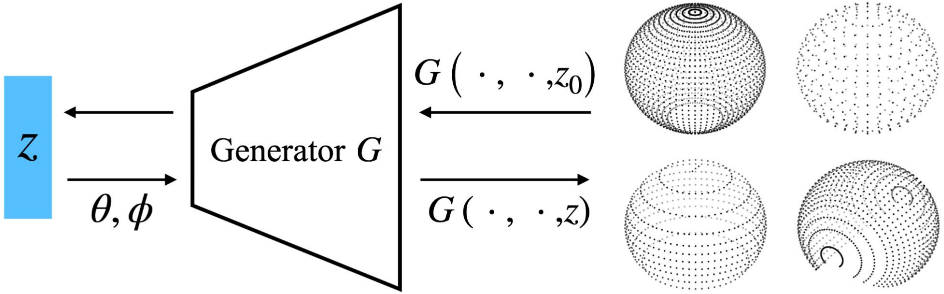


CIPIC

HRTF field for personalized HRTF modeling

We proposed ***HRTF field*** to alleviate the differences of spatial sampling schemes, enabling mix-database training [1].

The basic idea is to view an HRTF as a discrete sample from the underlying continuous-space HRTF and **model the continuous function** directly.

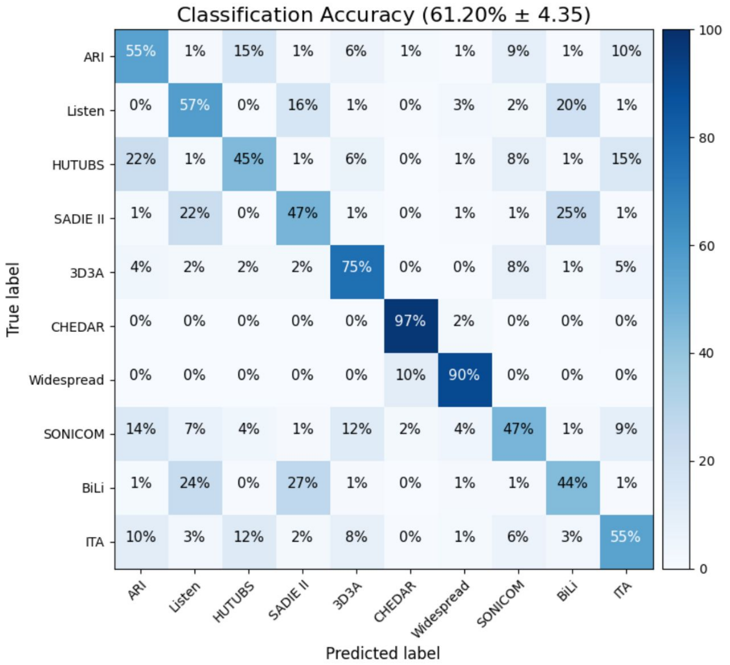


[1] Zhang, You, Yuxiang Wang, and Zhiyao Duan. "HRTF field: Unifying measured HRTF magnitude representation with neural fields." *ICASSP 2023*.

Differences beyond spatial sampling schemes

A recent study shows that there are other significant differences across HRTF databases.

This would hinder the training process.



Classification Accuracy (61.20% ± 4.35)

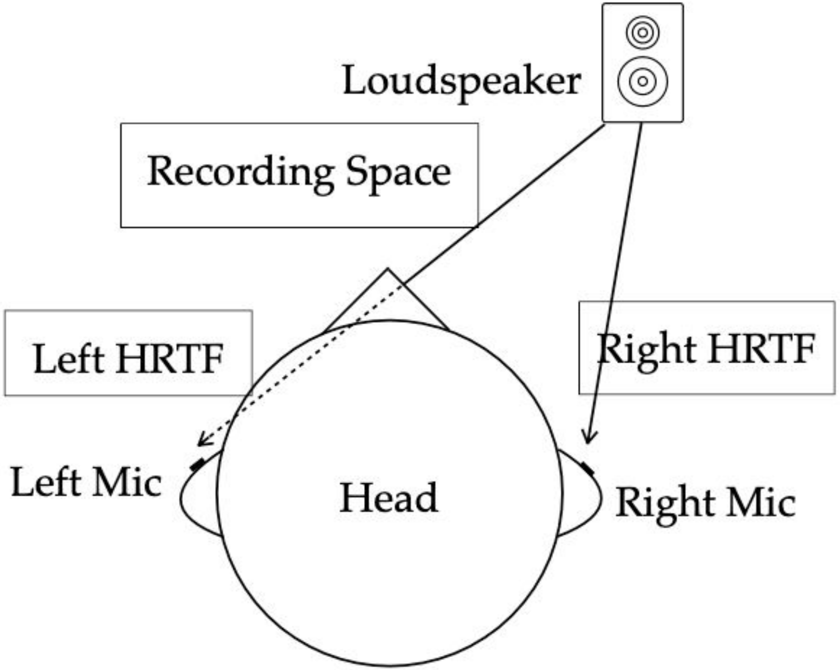
True label \ Predicted label	ARI	Listen	HUTUBS	SADIE II	3D3A	CHEDAR	Widespread	SONICOM	BiLi	ITA
ARI	55%	1%	15%	1%	6%	1%	1%	9%	1%	10%
Listen	0%	57%	0%	16%	1%	0%	3%	2%	20%	1%
HUTUBS	22%	1%	45%	1%	6%	0%	1%	8%	1%	15%
SADIE II	1%	22%	0%	47%	1%	0%	1%	1%	25%	1%
3D3A	4%	2%	2%	2%	75%	0%	0%	8%	1%	5%
CHEDAR	0%	0%	0%	0%	0%	97%	2%	0%	0%	0%
Widespread	0%	0%	0%	0%	0%	10%	90%	0%	0%	0%
SONICOM	14%	7%	4%	1%	12%	2%	4%	47%	1%	9%
BiLi	1%	24%	0%	27%	1%	0%	1%	1%	44%	1%
ITA	10%	3%	12%	2%	8%	0%	1%	6%	3%	55%

Pauwels, Johan, and Lorenzo Picinali. "On the relevance of the differences between HRTF measurement setups for machine learning." *ICASSP 2023*.

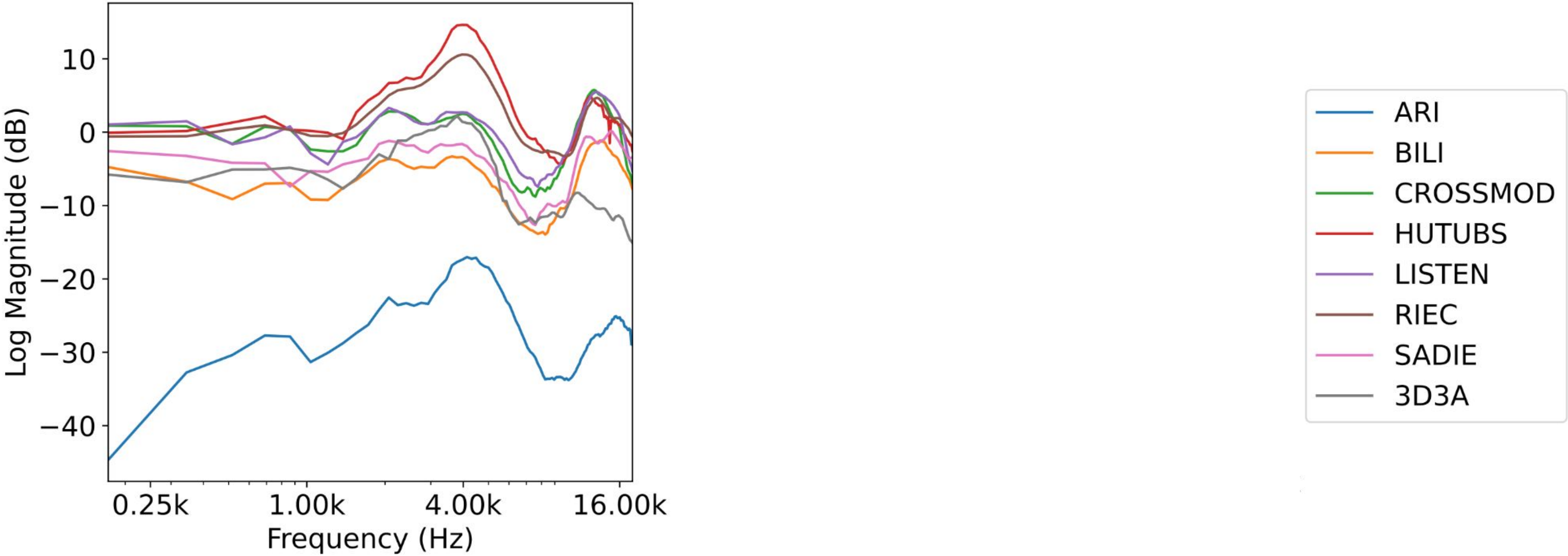
Investigating cross-database differences

Frequency response of:

1. Different loudspeakers,
2. Recording spaces,
3. Microphones.



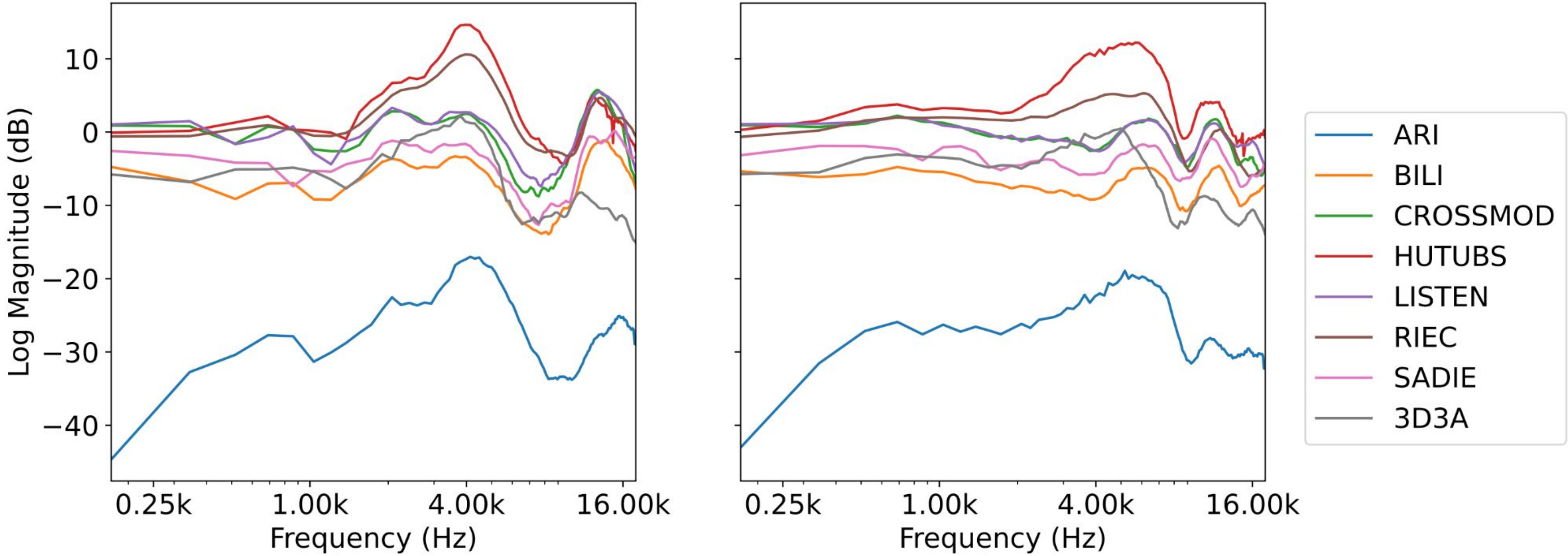
Average HRTFs across subjects



(a) at source position (0, 0)

There are systematic differences in measurement system responses at each source position.

Average HRTFs across subjects



(a) at source position (0, 0)

(b) at source position (90, 0)

These position-dependent systematic differences in measurement system responses need to be removed.

Our method to normalize HRTFs

Unnormalized HRTF magnitude

$$HRTF_{\mathrm{normalized}}(\theta, \phi) = \frac{Y(\theta, \phi)}{HRTF_{\mathrm{avg}}(\theta, \phi)}$$

azimuth

elevation

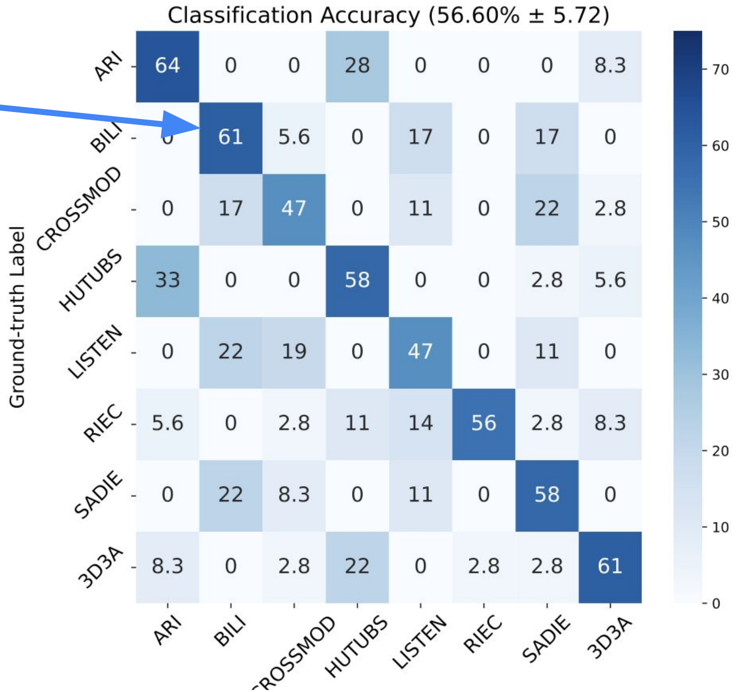
Average HRTF magnitude across subjects

HRTF database classification before normalization

Experimental setup:

- Frequency range
  - 200Hz to 18kHz
  - 104 frequency bins
- Total 144 subjects
  - 18 (the smallest size dataset) times 8
  - 432 HRTFs = 18 (subjects) * 12 (common positions) * 2 (ears)
- Model: kernel SVM

Numbers are percentage

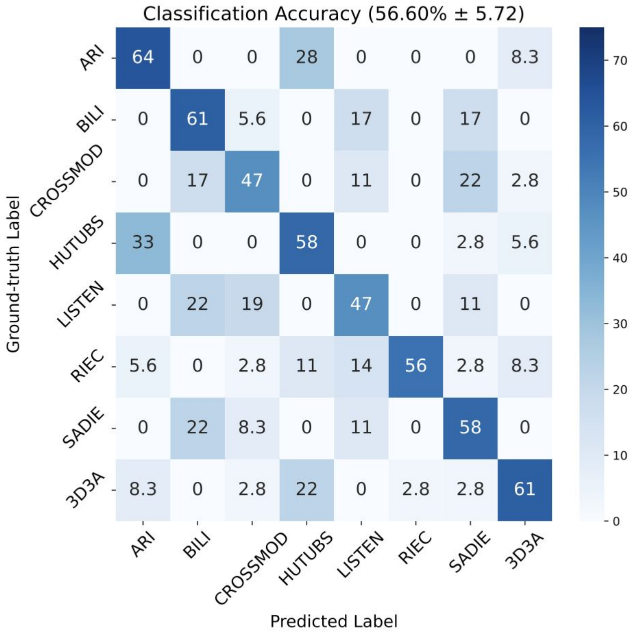


	ARI	BILI	CROSSMOD	HUTUBS	LISTEN	RIEC	SADIE	3D3A
ARI	64	0	0	28	0	0	0	8.3
BILI	0	61	5.6	0	17	0	17	0
CROSSMOD	0	17	47	0	11	0	22	2.8
HUTUBS	33	0	0	58	0	0	2.8	5.6
LISTEN	0	22	19	0	47	0	11	0
RIEC	5.6	0	2.8	11	14	56	2.8	8.3
SADIE	0	22	8.3	0	11	0	58	0
3D3A	8.3	0	2.8	22	0	2.8	2.8	61

SVM can easily tell where the HRTFs originate from.

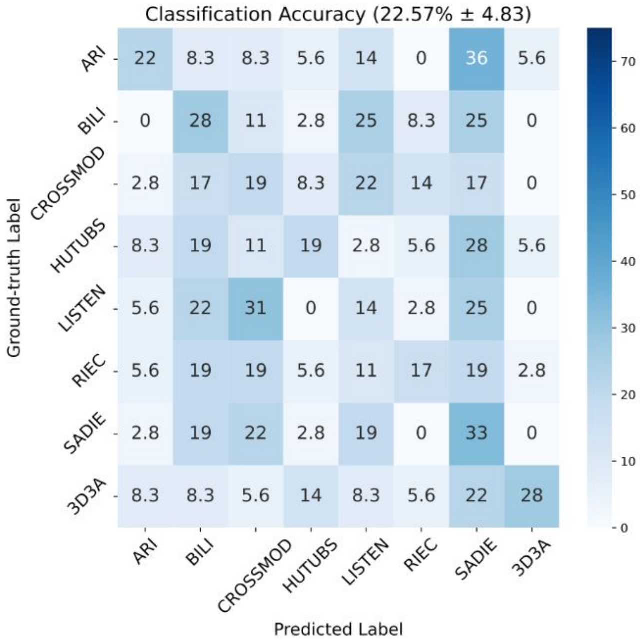
Pauwels, Johan, and Lorenzo Picinali. "On the relevance of the differences between HRTF measurement setups for machine learning." *ICASSP 2023*.

Our normalization successfully confuses SVM classifier


Classification Accuracy (56.60% ± 5.72)

	ARI	BILI	CROSSMOD	HUTUBS	LISTEN	RIEC	SADIE	3D3A
ARI	64	0	0	28	0	0	0	8.3
BILI	0	61	5.6	0	17	0	17	0
CROSSMOD	0	17	47	0	11	0	22	2.8
HUTUBS	33	0	0	58	0	0	2.8	5.6
LISTEN	0	22	19	0	47	0	11	0
RIEC	5.6	0	2.8	11	14	56	2.8	8.3
SADIE	0	22	8.3	0	11	0	58	0
3D3A	8.3	0	2.8	22	0	2.8	2.8	61

$$HRTF_{\text{normalized}}(\theta,\phi)=\frac{Y(\theta,\phi)}{HRTF_{\text{avg}}(\theta,\phi)},$$


Classification Accuracy (22.57% ± 4.83)

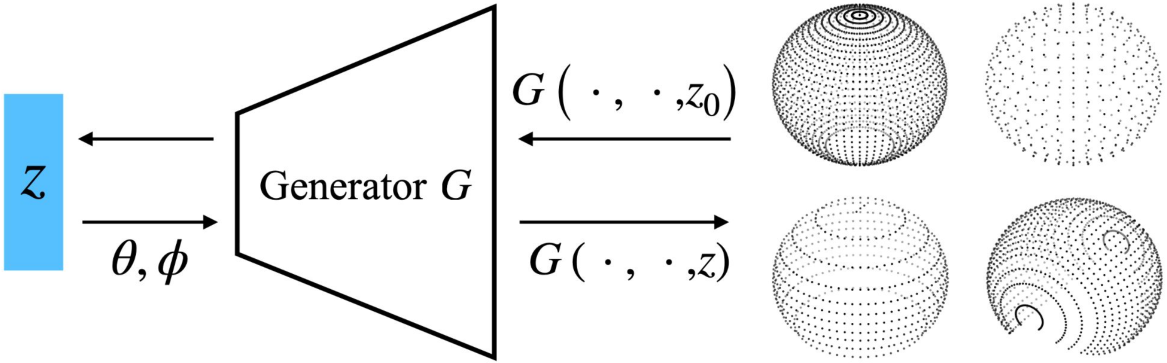
	ARI	BILI	CROSSMOD	HUTUBS	LISTEN	RIEC	SADIE	3D3A
ARI	22	8.3	8.3	5.6	14	0	36	5.6
BILI	0	28	11	2.8	25	8.3	25	0
CROSSMOD	2.8	17	19	8.3	22	14	17	0
HUTUBS	8.3	19	11	19	2.8	5.6	28	5.6
LISTEN	5.6	22	31	0	14	2.8	25	0
RIEC	5.6	19	19	5.6	11	17	19	2.8
SADIE	2.8	19	22	2.8	19	0	33	0
3D3A	8.3	8.3	5.6	14	8.3	5.6	22	28

Experiment with *HRTF field*

Use *HRTF field* to do cross-database reconstruction.

We train the generator with multiple databases combined.

Inference HRTFs from another database.



Zhang, You, Yuxiang Wang, and Zhiyao Duan. "HRTF field: Unifying measured HRTF magnitude representation with neural fields." *ICASSP 2023*.

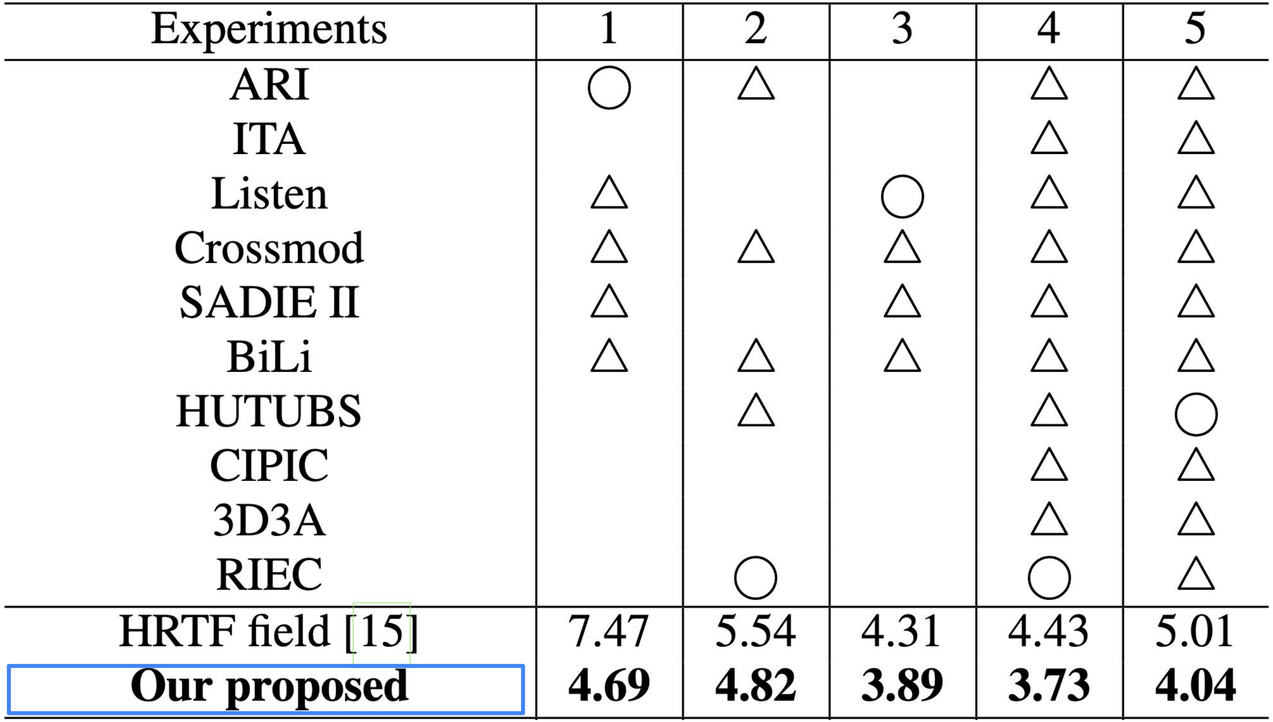
Evaluation on log-spectral distortion (LSD)

$$LSD(\overset{\text{Ground-truth}}{H}, \underset{\text{Predicted}}{H'}) = \sqrt{\frac{1}{\underset{\text{\# spatial locations}}{P}\underset{\text{\# frequency bins}}{N}}\sum_{\theta,\phi}\sum_{n}\left(20\log_{10}\left|\frac{H(\theta,\phi,n)}{H'(\theta,\phi,\underset{\text{Frequency index}}{n})}\right|\right)^{2}}$$

Our normalization improves cross-database reconstruction

Training sets are denoted with △

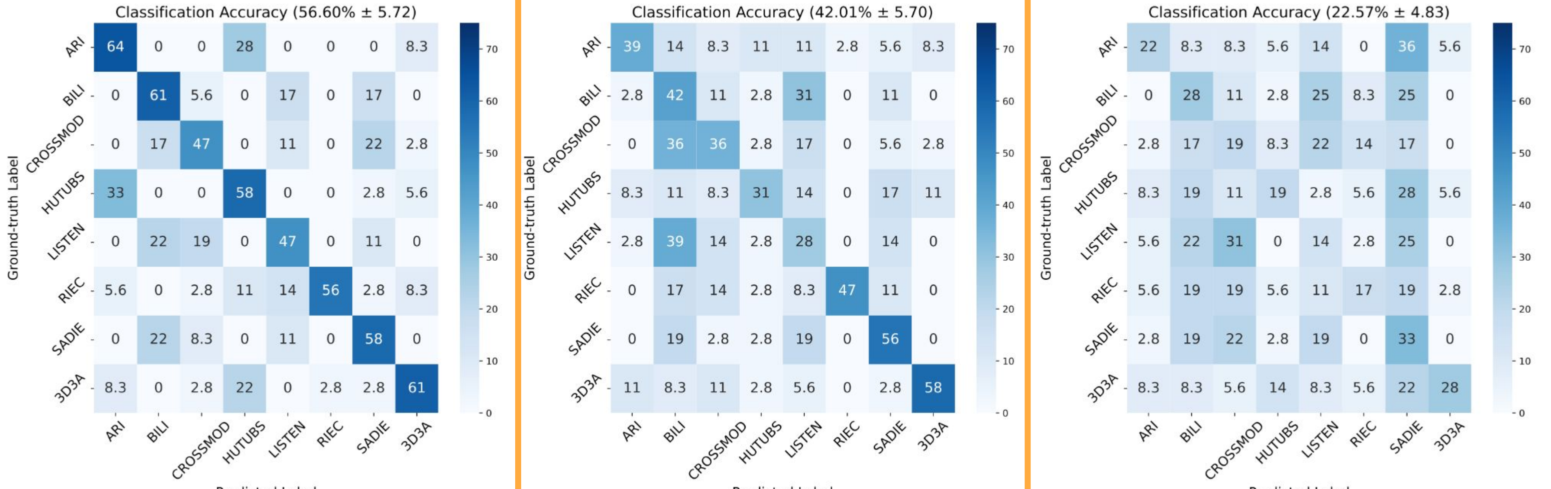
Testing sets are denoted with ○



Experiments	1	2	3	4	5
ARI	○	△		△	△
ITA				△	△
Listen	△		○	△	△
Crossmod	△	△	△	△	△
SADIE II	△		△	△	△
BiLi	△	△	△	△	△
HUTUBS		△		△	○
CIPIC				△	△
3D3A				△	△
RIEC		○		○	△
HRTF field [15]	7.47	5.54	4.31	4.43	5.01
Our proposed	**4.69**	**4.82**	**3.89**	**3.73**	**4.04**

HRTF field with our normalization method

Ablation study



(a) Unnormalized data

(b) Normalized by the average person HRTF derived from shared positions

(c) Normalized by the average person HRTF derived from individual position

HRTF field [15]	7.47	5.54	4.31	4.43	5.01
Our proposed	**4.69**	**4.82**	**3.89**	**3.73**	**4.04**
w/o position dependency	5.61	5.32	4.32	4.00	4.89
w/o ear dependency	5.11	5.11	3.98	3.94	4.67

Takeaways

- There are **position-dependent** systematic differences across HRTF databases.
- It is effective to normalize these differences using **average person HRTFs from individual positions.**
- Our proposed normalization method is promising to benefit many machine learning based methods for HRTF research.

ywen6@u.rochester.edu

Full text

Code